et le **partage de fichiers de poste-à-poste**. Avant de procéder à une présentation détaillée de ces trois systèmes, il serait judicieux de définir quelques notions-clé de ce domaine d'activités. On appelle **fournisseur de contenu** tout individu, association, entreprise ou autre institution impliqué(e) dans la diffusion de données (contenu) au moyen de l'internet. Le serveur d'origine d'un contenu particulier (tel qu'une page Web) est le serveur sur lequel l'objet a initialement été enregistré et qui continue d'héberger cet objet.

2.9.1 Mise en mémoire cache sur le Web

Un **cache Web**, également connu sous le nom de **serveur proxy** (mandataire), est une entité de réseau qui exécute des requêtes HTTP pour le compte d'un serveur d'origine. Le cache Web dispose de sa propre capacité de stockage, au sein de laquelle il conserve une copie d'objets récemment sollicités. Comme le montre la figure 2.27, un navigateur Web peut être configuré de manière que toutes les requêtes de son utilisateur soient préalablement dirigées vers un tel serveur (une procédure toute simple avec les navigateurs Netscape et Microsoft). Illustrons ce mode de fonctionnement par un exemple. Supposons qu'un navigateur sollicite l'objet `http://universitex.edu/campus.gif`. Voici les différentes étapes de la requête :

1. Le navigateur établit une connexion TCP avec le cache Web. Il lui envoie une requête HTTP concernant l'objet désiré sur le cache Web.
2. Le cache Web vérifie s'il dispose d'une copie de l'objet en local. Si c'est le cas, il transmet directement l'objet au navigateur client au sein d'un message de réponse HTTP.
3. Si le cache Web ne dispose pas de l'objet, il se connecte au serveur d'origine `www.universitex.edu`, auquel il envoie une requête HTTP concernant l'objet. Sur réception de la requête, le serveur d'origine transmet l'objet désiré au cache Web, par le biais d'un message de réponse HTTP.
4. Le cache Web fait une copie locale de l'objet reçu, puis l'envoie par un message HTTP au navigateur client (le long de la connexion TCP active entre le navigateur du client et le serveur Web).

Notez qu'un cache est simultanément client et serveur. Lorsqu'il reçoit des requêtes de la part d'un navigateur et qu'il lui envoie des messages de réponse, il agit en tant que serveur. Lorsqu'il envoie des requêtes et reçoit des réponses de la part d'un serveur d'origine, il agit en tant que client.

Un cache Web est généralement placé sous la responsabilité d'un fournisseur d'accès, qui en assure l'installation et la maintenance. Une université peut par exemple installer un cache au sein de son réseau local et configurer tous les navigateurs Web du campus pour qu'ils y soient automatiquement dirigés. De la même manière, un fournisseur d'accès résidentiel (tel qu'AOL) dispose de plusieurs caches au sein de son réseau et il a configuré le logiciel de navigation qu'il distribue à ses abonnés de manière à ce qu'il y fasse systématiquement appel.

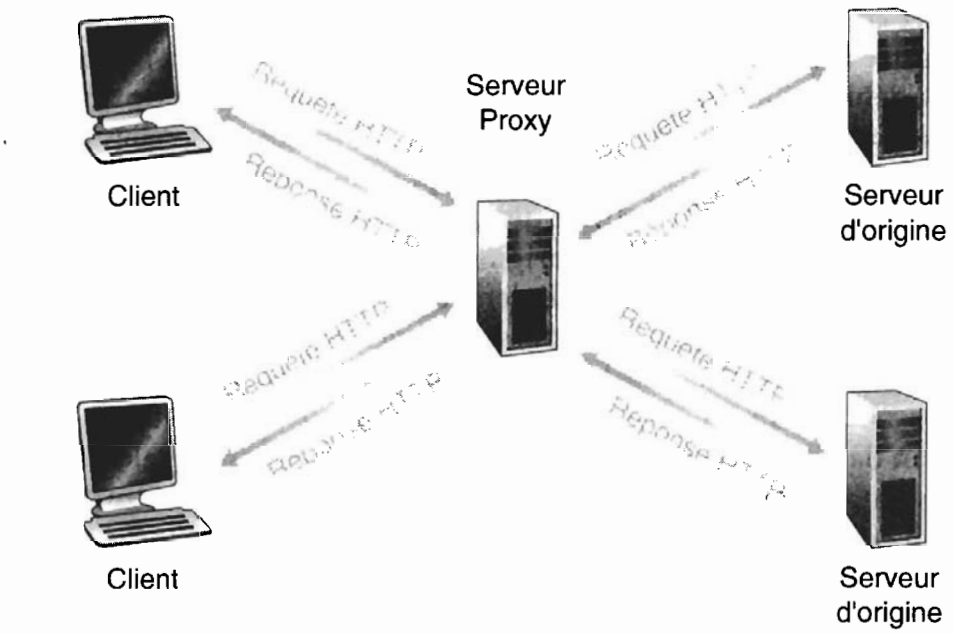

Figure 2.27 • Clients sollicitant des objets *via* un serveur cache.

La mise en mémoire cache sur le Web est considérée comme une forme de distribution de contenu dans la mesure où il s'agit de serveurs effectuant une réplication du contenu qui leur est soumis, tout en offrant aux utilisateurs le moyen d'y accéder. Le fournisseur de contenu ne supervise pas les opérations de réplication, qui sont exécutées en fonction des requêtes des utilisateurs.

La mise en mémoire cache sur le Web doit son succès à trois grands facteurs. Premièrement, cette technique est en mesure de réduire sensiblement le temps de réponse des différentes requêtes HTTP, notamment si le débit de la liaison entre le client et le serveur d'origine est beaucoup plus faible que celui de la liaison entre le client et le cache. Si ces deux derniers sont reliés par une connexion à haut débit, ce qui est souvent le cas, et si le cache dispose de l'objet désiré, le transfert se fera presque sans délai. Deuxièmement, comme on peut le voir à l'aide d'un exemple concret, les caches Web peuvent grandement contribuer à limiter la densité du trafic de la liaison d'accès internet d'une institution. Grâce à cette réduction de trafic, l'institution en question (par exemple, une entreprise ou une université) peut économiser d'importantes ressources en débit, ce qui se traduit par de sérieuses économies. Qui plus est, les serveurs cache contribuent à limiter le trafic du réseau internet en général, rendant par là un précieux service à l'ensemble des applications Web. Troisièmement, un réseau internet doté de nombreux caches Web, au niveau institutionnel, régional et national, constitue une infrastructure de distribution de contenu (CDN) plus rapide, y compris en ce qui concerne les fournisseurs de contenu ayant recours à des serveurs peu rapides et dotés de liaisons d'accès à faible débit. Si le contenu proposé par un tel prestataire devait par exemple acquérir une notoriété subite, ce contenu ferait rapidement l'objet d'une duplication au sein des nombreux caches internet et deviendrait par conséquent accessible au grand public sans aucun risque de saturation.

Afin de mieux comprendre les avantages associés aux serveurs cache, considérons l'exemple illustré à la figure 2.28, où sont représentés deux réseaux : un réseau local (LAN) à haut débit et l'internet, connectés l'un à l'autre au moyen de deux routeurs reliés par une liaison de 1,5 Mbit/s. Les serveurs d'origine sont reliés à l'internet, mais ils sont répartis dans le monde entier. Supposons que les objets aient une taille moyenne de 100 kbits et que le rythme des requêtes en provenance des navigateurs du réseau local auprès des serveurs d'origine soit de 15 requêtes par seconde. Supposons également que les messages de demande HTTP soient d'une taille négligeable, ne représentant donc aucun trafic supplémentaire dans les réseaux ou sur la liaison d'accès (entre routeur du LAN et routeur de l'internet). Imaginons enfin que le temps nécessaire au routeur de l'internet pour transmettre une requête HTTP dans un datagramme IP et pour recevoir une réponse généralement répartie sur plusieurs datagrammes IP soit de 2 secondes en moyenne. De manière informelle, ce délai est appelé « délai internet ».

Le temps de réponse total, soit le temps d'attente entre l'émission d'une requête et la réception de l'objet sollicité, est la somme du temps de transmission sur le réseau local, du temps d'accès (c'est-à-dire le temps d'acheminement entre les deux routeurs) et du transit sur l'internet. Essayons maintenant d'en faire une estimation très grossière : l'intensité du trafic sur le LAN (voir section 1.6) est de

(15 requêtes/s) • (100 kbit/requête)/(10 Mbit/s) = 0,15

tandis que l'intensité du trafic sur la liaison d'accès (du routeur internet au routeur du LAN) est de

(15 requêtes/s) • (100 kbit/requête)/(1,5 Mbit/s) = 1

Une intensité de trafic de 0,15 dans un LAN se traduit par un temps de réponse de l'ordre de la dizaine de millisecondes, soit une durée négligeable. Cependant, comme précisé à la section 1.6, plus cette valeur se rapproche de l'unité (comme c'est le cas pour la liaison d'accès de la figure 2.28), plus le temps de réponse peut prendre des proportions gênantes, allant jusqu'à compromettre tout transfert de données. Ainsi, le temps de réponse moyen d'une requête pourra atteindre plusieurs minutes, ce qui est bien au-delà d'une durée acceptable pour les utilisateurs du réseau. Il est donc impératif de trouver une solution.

Une première possibilité consisterait à augmenter le débit de la liaison d'accès de 1,5 Mbit/s actuels à 10 Mbit/s (par exemple). Ceci aurait pour effet de réduire l'intensité du trafic sur la liaison d'accès à une valeur de 0,15, ce qui, comme nous l'avons vu, correspond à un délai négligeable de routeur à routeur. Dans ce cas, le temps de réponse total se limiterait aux deux secondes imposées par le délai internet. Mais, malheureusement, cette solution implique de lourdes dépenses.

Une autre possibilité consiste à conserver la liaison d'accès actuelle et à installer un cache Web au sein du réseau d'entreprise. Cette solution est illustrée à la figure 2.29.

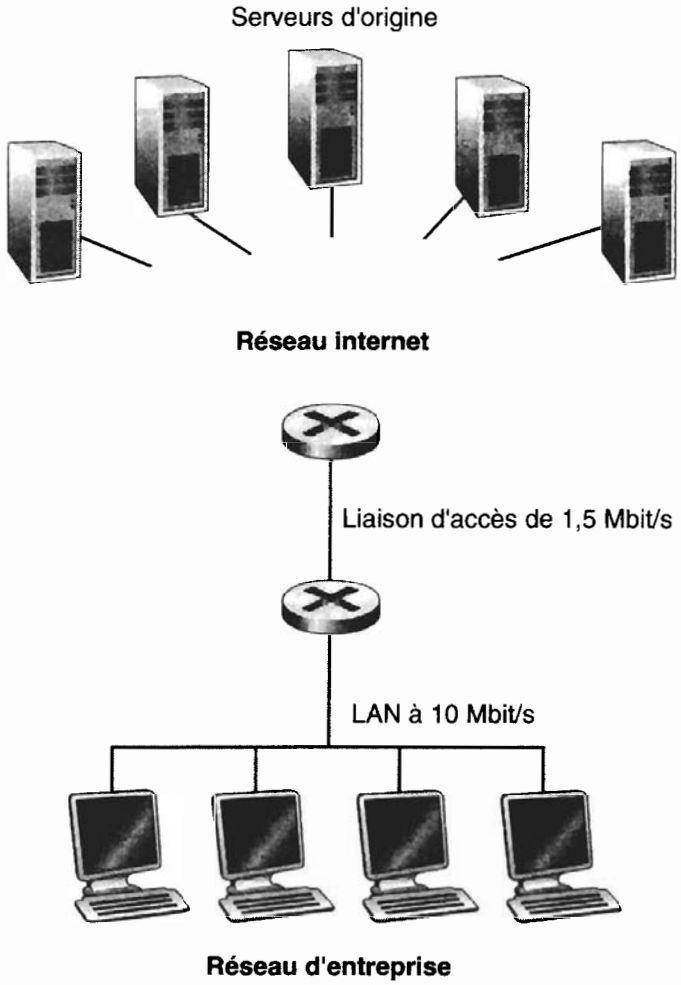

Figure 2.28 • Goulet d'étranglement entre un réseau d'entreprise et l'internet.

Dans la pratique, le taux de réussite, c'est-à-dire la proportion de requêtes satisfaites par un cache donné, varie généralement entre 0,2 et 0,7. Imaginons, à titre d'illustration, que le cache de notre réseau local présente un taux de réussite de 0,4. Étant donné que les clients et le cache sont connectés au même réseau local à haut débit, 40 % des requêtes seront satisfaites dans un laps de temps inférieur à 10 ms, soit presque immédiatement. Et malgré le fait que 60 % des requêtes soient toujours destinées aux serveurs d'origine, l'intensité du trafic sur la liaison d'accès sera réduite de 1,0 à 0,6 sachant qu'une valeur inférieure à 0,8 en présence d'une connexion à 1,5 Mbit/s correspond déjà à un temps de réponse très acceptable, de l'ordre de quelques dizaines de millisecondes à peine. Ce délai est donc négligeable par rapport aux deux secondes du délai internet calculées précédemment. Au vu de ces considérations, le temps de réponse moyen est de

$$0{,}4 \bullet (0{,}010\ \text{s}) + 0{,}6 \bullet (2{,}01\ \text{s})$$

soit à peine plus de 1,2 s. Cette seconde solution procure, à moindre frais, un temps de réponse encore plus faible que la première. La plupart des serveurs cache utilisant des logiciels du commerce, l'achat d'un cache Web s'impose comme la solution de loin la plus économique.

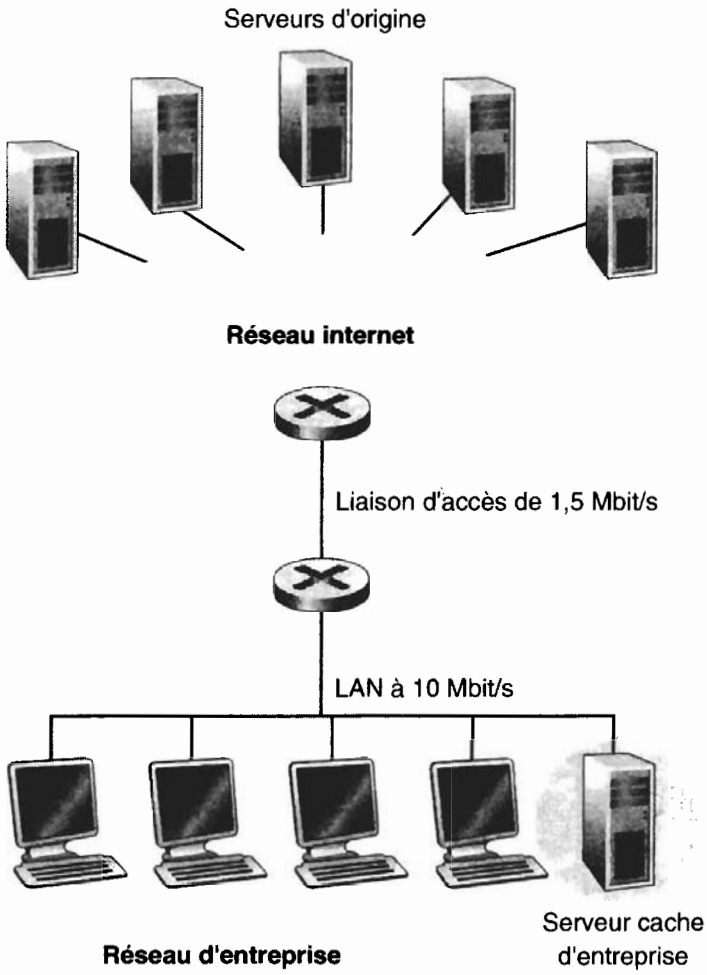

Figure 2.29 • Installation d'un serveur cache au sein d'un réseau d'entreprise.

Mise en mémoire cache coopérative

Afin d'améliorer leur efficacité, il est possible de faire coopérer plusieurs caches Web, même s'il s'agit de postes éloignés géographiquement. Un serveur cache d'entreprise peut par exemple être configuré de manière à envoyer ses requêtes HTTP à un cache placé au sein du réseau fédérateur d'un fournisseur d'accès. De cette manière, si le premier ne dispose pas de l'objet sollicité dans sa mémoire locale, il transmet la requête HTTP au second. Celui-ci extrait alors l'objet de sa propre mémoire ou du serveur d'origine, s'il ne dispose d'aucune copie lui non plus, puis il l'envoie au cache d'entreprise à l'intérieur d'un message de réponse HTTP, qui le transmet finalement à l'auteur de la requête. Lorsqu'un objet passe au travers d'un cache (d'entreprise ou fédérateur), ce dernier en garde une copie dans sa mémoire locale. Mais l'avantage de passer par un serveur cache de plus haut niveau, tel qu'un serveur cache fédérateur, permet de desservir un nombre d'utilisateurs plus important et présente de meilleurs taux de réussite.

Un exemple de système de mise en mémoire cache coopérative est illustré par Janet Web Cache Service (JWCS), qui est un cache coopératif national desservant les universités

britanniques [JWCS 2002]. Au sein du JWCS, plus de 200 serveurs cache transmettent des requêtes à un cache national traitant plus de 100 millions de transactions par jour. Ceux-ci se communiquent les objets les uns aux autres au moyen d'une combinaison entre les protocoles HTTP et ICP (*Internet Caching Protocol*). ICP est un protocole de niveau d'application qui permet à un cache d'en interroger un autre rapidement pour savoir s'il dispose d'un document déterminé [RFC 2186]. HTTP sert ensuite au transfert de l'objet entre le cache client et le cache serveur ainsi constitués. ICP est utilisé par la majeure partie des systèmes coopératifs et il est entièrement compatible avec Squid, un logiciel de mise en mémoire cache sur le Web de grande diffusion [Squid 2002]. Les lecteurs intéressés par le protocole ICP et ses applications peuvent consulter [Luotonen 1998 ; Ross 1998 ; Wessels 2001] ainsi que le RFC relatif à ICP [RFC 2186].

Un autre mode de mise en mémoire cache coopérative consiste en un regroupement de serveurs cache en grappe, souvent installés au sein du même réseau local. Un simple serveur cache est souvent remplacé par ce type de grappe lorsqu'il n'est plus en mesure de traiter tout le trafic du réseau ou d'assurer une capacité de stockage suffisante. Bien que le regroupement de serveurs cache constitue une manière naturelle d'adapter un réseau à l'augmentation de trafic, les structures en grappe ne manquent pas de poser un nouveau problème : lorsqu'un navigateur souhaite envoyer un objet spécifique, vers quel serveur de la grappe doit-il se tourner ? Ce problème peut être résolu à l'aide de la technique dite « d'acheminement par hachage » décrite au chapitre 7. Dans le plus simple des modes d'acheminement par hachage, le navigateur découpe l'URL, et selon le résultat, oriente son message de demande vers l'un ou l'autre des serveurs de la grappe. Si tous les navigateurs utilisent la même méthode de hachage, un objet ne peut jamais être présent simultanément dans plusieurs caches de la grappe, et si l'objet s'y trouve effectivement, le navigateur dirige toujours sa requête vers le bon cache. L'acheminement par hachage est l'essence même du protocole CARP (*Cache Array Routing Protocol*), utilisé par les systèmes de mise en mémoire cache de Netscape et de Microsoft. Vous en apprendrez davantage en essayant de résoudre les problèmes proposés à la fin de ce chapitre. Et si vous souhaitez en savoir encore plus sur cette technique d'acheminement ou sur le protocole CARP, référez-vous à [Ross 1997 ; Luotonen 1998 ; Wessels 2001].

2.9.2 Réseaux de distribution de contenu (CDN)

Les fournisseurs d'accès, qu'il s'agisse d'universités, d'entreprises ou de réseaux fédérateurs administrés par des fournisseurs spécialisés, s'équipent en caches Web afin d'offrir un meilleur service à leurs clients en aval (c'est-à-dire leurs utilisateurs réels et les ISP de niveau inférieur). Plus précisément, lorsqu'un fournisseur d'accès installe un cache, les clients en aval bénéficient théoriquement de meilleurs temps de réponse sur les pages Web les plus visitées.

Les **réseaux de distribution de contenu** (CDN, *Content Distribution Network*) reposent sur un « modèle commercial » différent de la mise en mémoire cache décrite précédemment. Les abonnés d'un tel réseau ne sont plus les fournisseurs d'accès, mais les fournisseurs de contenu. Un fournisseur de contenu (tel que Yahoo, par exemple)